

Generating random networks with given degree-degree correlations and degree-dependent clustering

Andreas Pusch, Sebastian Weber, and Markus Porto

Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr. 8, 64289 Darmstadt, Germany

(Dated: February 2, 2008)

Random networks are widely used to model complex networks and research their properties. In order to get a good approximation of complex networks encountered in various disciplines of science, the ability to tune various statistical properties of random networks is very important. In this manuscript we present an algorithm which is able to construct arbitrarily degree-degree correlated networks with adjustable degree-dependent clustering. We verify the algorithm by using empirical networks as input and describe additionally a simple way to fix a degree-dependent clustering function if degree-degree correlations are given.

PACS numbers: 89.75.Hc, 05.40.-a

I. INTRODUCTION

Modeling empirical networks as random networks is an important approach in the effort of studying topology and dynamics of complex networks. The first attempts in constructing random networks which exhibit some of the common features regularly found in empirical networks from fields as different as biology, social sciences, and technology have mostly aimed at understanding the origin of scale-free degree distributions (the degree of a vertex being its number of connections) and small average distances among vertices [1, 2]. However, it has been found that there are other important statistical quantities that profoundly influence the structure of complex networks and consequently the dynamics taking place on them. Notably among them are degree-degree correlations of vertices [3, 4, 5] and the abundance of motifs [6, 7, 8]. The smallest and probably most important motif in undirected graphs is the triangle. Its abundance is called clustering and several measures have been proposed to quantify it [9]. Some refined network growing mechanisms which extend the preferential attachment scheme introduced by Barabási and Albert to generate “scale-free” graphs [1] that are either correlated or clustered have been proposed [10, 11, 12, 13]. Those algorithms are, however, restricted in the correlation and clustering patterns they are able to produce.

Therefore, some efforts have recently been undertaken to overcome these restrictions. For example, the very successful configuration model (CM) algorithm [14, 15, 16], capable of generating random networks with an *a priori* given degree distribution, has been extended to include either degree-degree correlations or clustering properties of networks. Serrano and Boguñá presented an algorithm capable of tuning the degree-dependent clustering coefficient as well as the degree distribution [17]. Additionally, they pointed out that clustering and degree-degree correlation are deeply entwined, the latter limiting the former especially for vertices of high degree, in particular for disassortative networks where vertices of high degree are preferentially connected to vertices of low degree and vice versa. As both properties, clustering and correlations, are very important for the structure of a network and strongly related to each other, it is a natu-

ral ansatz to control degree-dependent clustering and the correlation pattern simultaneously to achieve better null models of complex networks. In this manuscript, we propose an algorithm to construct random networks with given degree-degree correlation structure and degree-dependent clustering. It is organized as follows: Section II introduces the network clustering and correlation measures used. Section III describes the algorithm to construct degree-degree correlated and clustered networks and verifies our scheme by applying it to empirical networks. Section IV presents a simple way to create networks with certain correlations and clustering and shows some results of this approach. In Section V we briefly summarize.

II. NETWORK CORRELATION AND CLUSTERING MEASURES

Two-point degree-degree correlations can statistically be described via a degree-degree correlation function $P(j, k)$ which is the probability that a randomly chosen edge has vertices of degrees j and k at its ends. In the case of uncorrelated networks, the correlation function factorizes into $P^u(j, k) = kP(k)jP(j)/\langle k \rangle^2$, where $P(k)$ is the degree distribution. Thus it appears natural to define a correlation function $f(j, k)$ as

$$f(j, k) = \frac{P(j, k)}{P^u(j, k)}. \quad (1)$$

Values of $f(j, k)$ different from 1 signal degree-degree correlations in the underlying network. A simpler but more coarse-grained manner to quantify degree-degree correlations is the average nearest neighbor function $k_{nn}(k)$, describing the average degree of neighbors of vertices with degree k . It can be calculated from the conditional probability $P(j|k) = P(j, k) \langle k \rangle / [kP(k)]$ as

$$k_{nn}(k) = \sum_j jP(j|k). \quad (2)$$

A network with an (de-)increasing $k_{nn}(k)$ is called (dis-)assortatively correlated.

Clustering was originally defined by Watts and Strogatz [2] for the vertex i to be

$$c_i = \frac{2T_i}{k_i(k_i - 1)}, \quad (3)$$

where T_i denotes the number of triangles passing through vertex i . Clearly this measure is a three-point dependent value as the number of triangles requires knowledge over three connected vertices at the same time. However, it is common use to average the clustering coefficients c_i of all vertices with the same degree k together, yielding a degree-dependent clustering coefficient

$$c(k) = \frac{1}{k(k-1)P(k)N} \sum_{i \in \Upsilon(k)} 2T_i, \quad (4)$$

where $\Upsilon(k)$ denotes the set of vertices with degree k .

Serrano and Boguñá pointed out that the degree-dependent clustering $c(k)$ is restricted by degree-degree correlations and is often found to be a decreasing function of k [17]. They calculated an upper limit $\lambda(k)$ of $c(k)$ dependent on the degree-degree correlation function $P(j, k)$. The main reasoning is that an edge cannot be part of more triangles than $\min(k_i, k_j) - 1$ with k_i and k_j being the degrees of the vertices connected by it. This results in a constraint on the number of triangles T_i for any vertex i ,

$$T_i \leq \sum_j a_{ij} [\min(k_i, k_j) - 1]. \quad (5)$$

Here a_{ij} is the network's adjacency matrix. The upper limit $\lambda(k)$ of the degree-dependent clustering $c(k)$ can then be written as

$$\lambda(k) \equiv 1 - \frac{1}{k-1} \sum_{j=1}^k (k-j)P(j|k) \geq c(k). \quad (6)$$

This function is always a decreasing function of k and its slope depends strongly on the average neighbor degree $k_{\text{nn}}(k)$. This means that degree-dependent clustering $c(k)$ can be written as

$$c(k) = c_{\text{eff}}(k)\lambda(k) \quad (7)$$

with $0 \leq c_{\text{eff}}(k) \leq 1 \forall k$. Thus $c_{\text{eff}}(k)$ can be regarded as an effective degree-dependent clustering, once degree-degree correlations are fixed.

In the following, we describe an algorithm that is able to control the two quantities $P(j, k)$ and $c(k)$ (or $c_{\text{eff}}(k)$) simultaneously.

III. ALGORITHM

As already stated, there exists an algorithm to create networks with a given degree distribution and a given level of clustering published by Serrano and Boguñá [17]. We incorporated some of their basic ideas into our approach which additionally fixes the degree-degree correlations besides the degree-dependent clustering.

The overall scheme of the algorithm to construct a network with N vertices and a given $P_d(j, k)$ and $c_d(k)$, $P_d(j, k)$ being the number of connections between vertices with degrees k and j (double that number if $k = j$), and $c_d(k)$ being the number of triangle edges constituted by vertices with degree k , is the following:

We begin by assigning a number of stubs (the target degree) to every vertex according to the degree distribution $P_d(k)$, which is calculated from $P_d(j, k)$ as $P_d(k) = \sum_j P_d(j, k)/k$.

The next step is to get a list of degrees of triangle-corners, which shall contain $c_d(k)$ entries with value k . We also get a copy P'_d of $P_d(j, k)$ and c'_d of $c_d(k)$, which are dynamical quantities in the sense that these shall be decreased with every connection and triangle build. Thus for every connection built we decrease the appropriate entry in the $P'_d(j, k)$ matrix by 1 and for every triangle built (for every connection we place, we check for simultaneous neighbors of the involved vertices as any shared neighbor accounts for a new triangle built) to delete one entry from the triangle list and to decrease $c'_d(k)$ by 1 for every degree involved.

Then we start to build all triangles in the triangle list one by one. Let v_i be the vertices involved and k_i their target degree.

1. We draw a random entry k_1 from the triangle list and draw a corresponding vertex v_1 with at least one free stub. If we cannot find such a vertex, we delete all entries with value k_1 from the triangle list and start again.
2. Now, we choose with uniform probability either (a) an edge or (b) a stub of vertex v_1 out of a list created by omitting all edges for whose end vertex no more triangles can be build (i.e. $c'_d(k) = 0$). In case of (a), we have chosen an edge and the end vertex is v_2 . If we have drawn a stub (b), we get a vertex v_2 in the same manner as we got vertex v_1 with the further condition $P'_d(k_1, k_2) > 0$. If it is not possible to find a k_2 fulfilling this condition, we delete all entries with value k_1 from the triangle list and start again.
3. Next, we draw (a) an edge or (b) a stub of vertex v_2 from a list like we did in the preceding step for vertex v_1 , but with edges inserted into the list only if they are fulfilling the supplementary condition of $P'_d(k_1, k_3) > 0$ or vertex v_3 being connected to vertex v_1 and vertices v_1, v_2 , and v_3 not already constituting a triangle. Having drawn an edge (a), we close the triangle by adding the missing edges and updating all dynamic quantities. Having drawn a stub (b), we choose a k_3 from the triangle list consistent with k_1 and k_2 . It might happen that this is not possible and we start again. When we got a k_3 , we draw a vertex v_3 which either has enough free stubs or is already connected to vertex v_1 or v_2 , add the missing edges, and update all dynamic quantities.

Note that in steps 2 and 3 the case of two or three degrees being the same has to be properly taken into account in

order not to build too much triangles or connections, and that self-connections are forbidden.

Those steps are repeated until we cannot build any triangles anymore. This point may be defined by a maximum number of successive tries that did not result in a triangle being built or until the triangle list is empty.

Afterwards we build the rest of the graph by randomly choosing edges out of the remaining edge list, which contains $P'_d(k_1, k_2)$ entries (k_1, k_2) for all degrees k_1, k_2 . We choose randomly two non-identical vertices with stubs left and build the edges (if the vertices are not already connected) and delete the edge we chose from the edge list. We repeat this until the edge list is empty or we cannot find any vertices which still lack connections and are not already connected to each other. If there are edges we could not build (typically there is no edge left, and very seldomly there are more than one or two edges left), we substitute them by randomly connecting vertices.

To validate our algorithm, we use two empirical networks as test cases: (i) the yeast protein-interaction network (PIN) constituent of 1,846 proteins [18] downloaded from Barabási's web site <http://www.nd.edu/~networks/>, (ii) a subset of the internet on the autonomous system level (AS) with 10,515 vertices (snapshot taken on 03/16/2001) taken from <http://www.cosin.org/>. All self- and multiple-edges were removed from each network. To test the validity of the algorithm, one measures the joint degree distribution $P(j, k)$ and the degree-dependent clustering $c(k)$ of the empirical networks and uses these functions as input for the construction algorithm. The resulting random network has to display the same joint degree distribution $P(j, k)$ (this implies that the degree distribution $P(k)$ is met as well) and the same degree-dependent clustering $c(k)$ as the empirical one. A very sensitive test to validate if the correlation structure of the reference and the random network indeed match is on the level of the correlation function $f(j, k)$, which varies on a much smaller scale than the joint degree distribution $P(j, k)$. Thus, comparing the reference correlation function $f_{\text{ref}}(j, k)$ with the resulting correlation function $f(j, k)$ by use of a correlation coefficient (1 means total agreement, -1 indicates that the two functions are of opposite sign and 0 means no correlation among the two functions in comparison) reveals almost complete agreement of (i) 0.9999(9) and (ii) 0.999(7). A density plot of the reference function versus the resulting correlation function in Fig. 1 verifies the excellent agreement of the correlation functions $f(j, k)$ and $f_{\text{ref}}(j, k)$, as the density of points is almost solely centered at the diagonal. The statistics per curve are 10^3 realizations for the AS network and 10^4 for the PIN.

However, the main and new point of our algorithm is its ability to conserve the degree-dependent clustering as well. The quality of agreement is shown in Fig. 2. We show a comparison between the degree-dependent clustering $c(k)$ of empirical and generated networks. One can see that the level of clustering in the PIN and AS network is well reproduced.

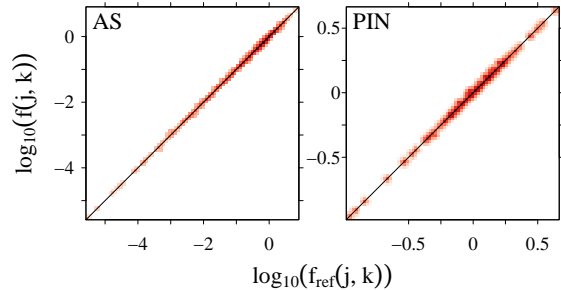


FIG. 1: (color online) Plot of the correlation function $f(j, k)$ of the random networks generated by the present algorithm vs the correlation function $f_{\text{ref}}(j, k)$ of the corresponding empirical network. The data is presented as a density plot. Darker red regions contain a higher density of data points, while brighter red indicates a lower density. A reference line $y = x$ is drawn as a guide to the eye.

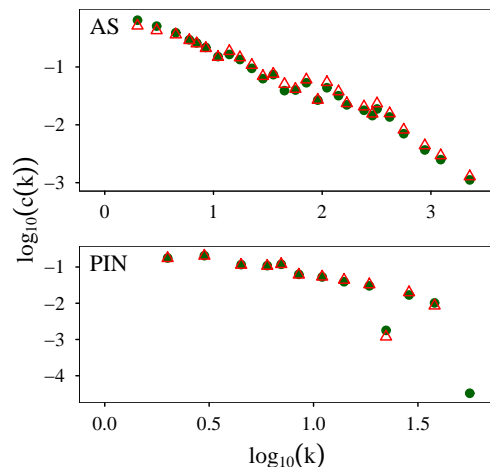


FIG. 2: (color online) Degree-dependent clustering coefficient $c(k)$ vs k of empirical graphs (open, red triangles) compared to their randomized versions generated by the present algorithm (full, green circles).

IV. CORRELATIONS AND CLUSTERING

We wish not only to be able to reproduce correlations and clustering in empirical networks, but also to create graphs from scratch that follow an adjustable correlation pattern expressed by the average nearest neighbor function $k_{\text{nn}}(k)$ and a tunable degree-dependent clustering coefficient $c(k)$. Eq. (7) defines $c_{\text{eff}}(k)$ as an effective clustering. So we might consider a graph showing

$$c_{\text{eff}}(k) = \mu, \quad (8)$$

with μ being a constant between 0 and 1, as an equally clustered graph throughout all degree classes. Therefore we may tune the level of clustering by changing μ . As we are able to control degree-degree correlations by use of the algorithm presented in [19], we can calculate the upper limit $\lambda(k)$ and therefore the degree-dependent clustering $c(k)$ from $P(j, k)$ and the target clustering $c_{\text{eff}}(k)$ via Eqs. (6) and (7). To get a discrete correlation func-

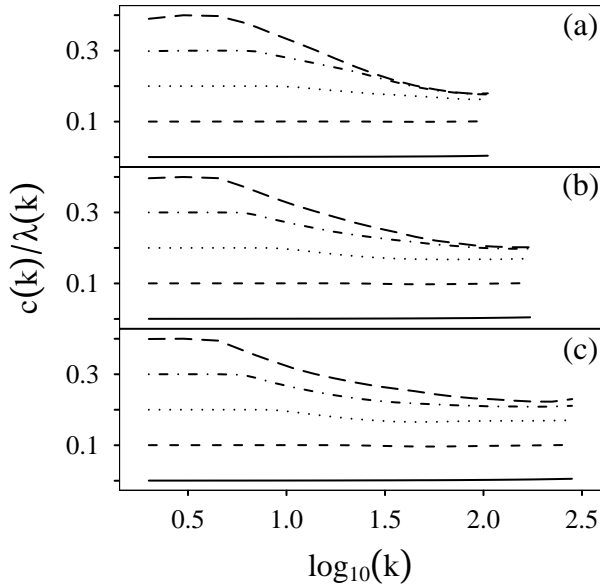


FIG. 3: Clustering $c_{\text{eff}}(k) = c(k)/\lambda(k)$ vs k exemplified for $N = 10^5$, $\gamma = 2.8$, $k_{\text{nn}}(k)$ as given in the main text, and $\mu = 0, 0.1, 0.2, 0.3$, and 0.4 (from bottom to top), (a) $\alpha = 0.2$ (assortative), (b) $\alpha = 0$, and (c) $\alpha = -0.2$ (disassortative).

tion $P_d(j, k)$, we first create a graph with a given degree distribution $P(k)$ and a correlation structure characterized by a given $k_{\text{nn}}(k)$ using the method presented in [19], and obtain its discrete correlation function $P_d(j, k)$. With Eq. (7) we get $c(k)$ and therefore the number of triangles per degree k as

$$c_d(k) = c(k)P(k)(k-1)N. \quad (9)$$

With $c_d(k)$, $P_d(j, k)$ and the resulting discretized $P_d(k)$ we have the input needed for our algorithm. To validate the algorithm we tested it for a scale-free graph ($P(k) \propto k^{-\gamma}$) with several levels of clustering and several degrees of assortativity using $k_{\text{nn}}(k) \propto \exp\left((\ln(1 + \frac{k}{k_{\text{min}}}))^\alpha\right)$ as

an example. The graph size has been set to $N = 10^5$ vertices. In order to avoid intrinsic degree-degree correlations caused by the constraint of no self- and multiple connections [16, 20], one has to limit the maximum degree to a k_{max} depending on the scale-free exponent γ and the level of (dis-)assortativity controlled by α [19].

In Fig. 3 we show the resulting $c_{\text{eff}}(k)$. The statistics per curve are 100 realizations each, with $P_d(j, k)$ drawn for each realization separately. One observes that a level of clustering close to $\mu = 1$ is not achievable with our algorithm. Low levels of clustering are very well reproducible, and medium levels of clustering are very well reproducible for lower degrees, but the higher the degree the more difficult it gets to cross a certain level of clustering, this level being dependent of the level of clustering of the lower degree classes. This behavior is not surprising as in calculating the upper bound $\lambda(k)$ it is assumed that all vertices i with a degree k' smaller than k have a clustering coefficient $c_i = 1$. Thus restrictions on the level of clustering of low degree vertices imply stronger restrictions on the level of clustering of high degree vertices. As changing the assortativity via α has only a minor effect on the effective clustering $c_{\text{eff}}(k)$ which can be reached, it seems that the effects of degree-degree correlations on clustering are well described by the upper bound $\lambda(k)$.

V. CONCLUSION

In summary, we have presented an algorithm which generates networks with an *a priori* fixed degree-degree correlation structure defined by the joint degree distribution $P(j, k)$ and an adjustable level of clustering defined by the degree-dependent clustering coefficient $c(k)$. As clustering and degree-degree correlations are suspected to play an important role in many dynamical processes taking place on networks, our algorithm may provide a very useful tool to systematically research the influences of those topological properties on different dynamics.

-
- [1] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
 - [2] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
 - [3] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. E **65**, 066130 (2002).
 - [4] M. E. J. Newman, SIAM Review **45**, 167 (2003).
 - [5] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
 - [6] A.-L. Barabási and E. Ravasz, Phys. Rev. E **67**, 026112 (2003).
 - [7] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
 - [8] S. Itzkovitz and U. Alon, Phys. Rev. E **71**, 026117 (2005).
 - [9] M. A. Serrano and M. Boguñá, Phys. Rev. E **74**, 056114 (2007).
 - [10] P. Holme and B. J. Kim, Phys. Rev. E **65**, 026107 (2002).
 - [11] A. Vázquez, Phys. Rev. E **67**, 056104 (2003).
 - [12] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
 - [13] A. Barrat and R. Pastor-Satorras, Phys. Rev. E **71**, 036127 (2005).
 - [14] M. Molloy and B. Reed, Random Structure and Algorithms **6**, 161 (1995).
 - [15] M. Molloy and B. Reed, Combinatorics, Probability and Computing **7**, 295 (1998).
 - [16] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras, Phys. Rev. E **71**, 027103 (2005).
 - [17] M. A. Serrano and M. Boguñá, Phys. Rev. E **72**, 036133 (2005).
 - [18] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature **411**, 41 (2001).
 - [19] S. Weber and M. Porto, Phys. Rev. E (in print) (2007), see <http://arxiv.org/abs/arXiv:0708.4161>.
 - [20] S. Dorogovtsev, J. Mendes, A. Povolotsky, and A. Samukhin, Phys. Rev. Lett. **95**, 195701 (2005).